Experiments on Visual Information Extraction with the Faces of Wikipedia

Md. Kamrul Hasan and Christopher Pal
Département de génie informatique et génie logiciel, Polytechnique Montréal
2500, Chemin de Polytechnique, Université de Montréal, Montrèal, Québec, Canada

Abstract

We present a series of visual information extraction experiments using the Faces of Wikipedia database - a new resource that we release into the public domain for both recognition and extraction research containing over 50,000 identities and 60,000 disambiguated images of faces. We compare different techniques for automatically extracting the faces corresponding to the subject of a Wikipedia biography within the images appearing on the page. Our top performing approach is based on probabilistic graphical models and uses the text of Wikipedia pages, similarities of faces as well as various other features of the document, meta-data and image files. Our method resolves the problem jointly for all detected faces on a page. While our experiments focus on extracting faces from Wikipedia biographies, our approach is easily adapted to other types of documents and multiple documents. We focus on Wikipedia because the content is a Creative Commons resource and we provide our database to the community including registered faces, hand labeled and automated disambiguations, processed captions, meta data and evaluation protocols. Our best probabilistic extraction pipeline yields an expected average accuracy of 77% compared to image only and text only baselines which yield 63% and 66% respectively.

Introduction

Wikipedia is one of the largest and most diverse encyclopedia in human history. There are about 550,000 biographies in the English version of Wikipedia and they account for about 15% of the encyclopedia (Kittur, Chi, and Suh 2009). This web-encyclopedia is constantly growing and being updated with new biographies, textual content and facial images. Furthermore, the presence of a Wikipedia biography page containing a photograph implies that the subject of the biography already has a public profile and the Wikipedia organization has mechanisms in place to resolve issues related to accuracy, privacy and the rights related to images. For example, most images are associated with meta-data explicitly indicating if imagery is officially in the public domain or has been given a creative commons designation. For these and many other reasons, these biography pages provide an excellent source of raw data to explore data mining algorithms and to produce a "big data" resource for computer vision experiments involving faces. Wikipedia also has a rich category structure that encodes many interesting semantic relationships between pages. We use the biographies in the Living People category from which we obtain 64,291 biography pages containing at least one detected face of a minimum resolution of 40×40 pixels.

Our mining goal here is to classify all faces detected within the images of a Wikipedia biography page as either positive or negative examples of the subject of the biography. While our technique could be easily extended to include images extracted from other web pages, we keep our work here focused on Wikipedia to both limit the scope of this paper and because of the numerous advantages of Wikipedia discussed both above and below. We are interested in particular in extracting faces automatically without using any prior reference face information. Indeed part of our motivation is that one could use Wikipedia as the starting point to automatically ramp up a larger web scale mining effort - for a search engine for example. Our overall approach is motivated by the desire to create a principled approach to manage uncertainty arising from different aspects of the extraction process. As such, we take the approach of dynamically constructing Bayesian networks and performing inference in these networks so as to correctly identify the true examples of a given person's face.

One of the many advantages of Wikipedia's biography page format is that the simple existence of a biography page for a given person typically implies that faces on the page are likely to be the person of interest. Biography pages with a single face detected in the image contain a face of the person of interest 93% of the time in our initial sampling and analysis. For multi face biography pages the problem is more challenging. In both cases we shall use information from many sources including image file names and various other sources of meta-data. Using various NLP techniques, we can define features that will help us to resolve many ambiguities; however, the fact that we have multiple faces detected in multiple images allows us to also combine NLP techniques with an approach to visual co-reference into one coherent model.

In addition to the creation and release of this Wikipedia derived dataset - including a large quantity of human labeled identity ground truth for facial images, another key contribution of our work here is the exploration, analyses and comparisons of different models and visual information ex-

traction strategies. In particular we present a novel approach to visual information extraction based on dynamically instantiated probabilistic graphical models. We also examine the importance of high quality image registration and compare our probabilistically formulated extraction process with a variety of more heuristic extraction approaches and baselines. Given the importance of visual comparisons for face extraction, along the way to formulating a principled solution to the visual extraction problem we have also developed a state of the art face verification technique.

Related Work

The 80-million tiny images (Torralba, Fergus, and Freeman 2008) and ImageNet (Deng et al. 2009) projects along with their associated evaluations are well known and are widely used for scene and object recognition research. The human face might be considered as a special type of object that has been studied intensely because of its importance. In recent years, facial analysis research attention has shifted towards the task of face verification and recognition in the wild - natural settings with uncontrolled illumination and variable camera positioning that is reflective of the types of photographs one normally associates with consumer, broadcast and press photos containing faces. Table 1 summarizes a number of prominent 'in the wild' face recognition databases and compares some of their key attributes with the dataset used in our work here which we refer to as the Faces of Wikipedia. Chokepoint collects imagery from a security camera (Wong et al. 2011). In contrast, the other databases use imagery from the Internet except for the Toronto Face Database (TFD) which consists of a collection of 30 pre-existing face databases, most of which were in fact collected under different controlled settings.

The grouping or clustering of faces in multiple images has been explored in a variety of contexts. Some prior work examining related but different situations include that of Zhang et al. (2004) where they used a visual similarity based optimization technique to group faces for a person in family albums. Anguelov et al. (2007) proposed a Markov Random Field model to disambiguate faces in personal photo albums in which they also used features derived from the clothing that people were wearing. Our work has some similarities to these types of applications but faces found in Wikipedia biographies have many additional types of information that can be used to solve our visual extraction problem.

The Labeled Faces in the Wild (LFW) is of particular interest to our work here as it has a large number of identities collected from the so called in the wild imagery. The underlying faces in the LFW were initially collected from press photos as discussed in Berg et al. (2004a). The original "Names and faces in the News" project (Berg et al. 2004b) sought to automate the process of extracting faces from press photos and their captions using both Natural Language Processing (NLP) and vision techniques. They used a per name clustering technique to associate a person's name and their face. In comparison, Guillaumin et al. (2012) proposes a metric learning technique for resolving the name and face association problem in the press photo data of Berg et al. (2004b). Our work here is similar in spirit, but our mining task is different in a number of respects. We outline a few of the key differences here. Firstly, the text captioning of Wikipedia images is not as standardized as the press photo captions that were used in Berg et al. (2004b). In contrast, Wikipedia does not strictly impose a particular format for the descriptive text of captions so the text is less structured than many news photo annotations. As such Wikipedia captions exhibit variability much more characteristic of what one might call "captions in the wild". Secondly, Wikipedia pages themselves are structured documents with various other useful clues concerning the underlying content of images. Images often have detailed comments in their meta-data and extremely long filenames using natural language to describe content. Third, we wish to resolve all the faces detected across all images from a Wikipedia biography page. As we shall see, we are able to exploit these aspect of the Wikipedia biography face mining problem to further increase extraction performance.

Name	Identities	Faces
TFD(1)[Suskind et al. (2010)]	963	3,874
Caltech 10k [Angelova et al. (2005)]	undefined	10,524
ChokePoint [Wong et al. (2011)]	29	64,204
YouTube Faces(2) [Wolf et al. (2011)]	1595	-
Face Tracer(3) [Kumar et al.(2008)]	undefined	17,000
PubFig(4) [Kumar et al. (2009)]	200	59,476
LFW [Huang et al. (2007)]	5,749	13,233
LFW (≥ 2)	1,680	9,164
The Faces of Wikipedia v.1	1,534	3,466
≥ 2 (currently labeled)	894	2,826
The Faces of Wikipedia v.2	59,000	68,000
≥ 2 (estimated, approx.)	9,000	18,000

(1) Also possess 112,234 unlabeled faces. (2) Consists of 3425 videos; no statics of faces was provided. (3) They possess a much larger database of 3.1 million faces; however, only 17,000 image http links are published. (4) Only image http links are provided.

Table 1: Some important 'in the wild' face databases

Our Extraction Technique

We present a high level view of our technique using a concrete example for the two images in Figure 1 found within the biography of Richard Parks [1]. Therein we outline the major sub-components of our overall system. We give more details further on in this paper. For a given biography, our mining technique dynamically creates probabilistic models to disambiguate the faces that correspond to the subject of the biography. These models integrate uncertain information extracted throughout a document arising from three different modalities: text, meta data and images. We also show an instance of our mining model for Mr. Parks in Figure 1. The image on the far left was contained in a Wikipedia infobox which is sometimes but not always found on the far right of a biography page. The second image was found in the body text of the biography. The model is a Bayesian network and can be used as a guide to our approach. Text and meta-data

[1] http://en.wikipedia.org/wiki/Richard_Parks (September, 2011)

features are taken as input to the bottom layer of random variables $\{X\}$, which influence binary (target or not target) indicator variables $\{Y\}$ for each detected face. The result of visual comparisons between all faces, detected in different images, are encoded in the variables $\{D\}$. Soft constraints are captured by the arcs and variables $\{S\}$.

Caption text: Richard Parks celebrating the end of the 737 Challenge at the National Assembly for Wales on 19 July 2011

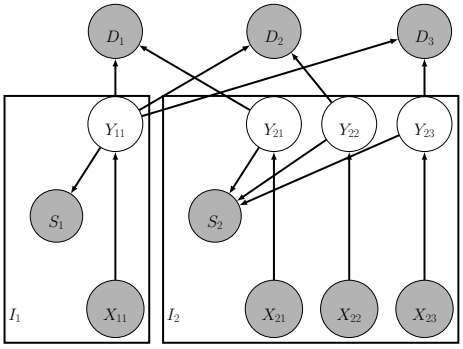

Figure 1: An instance of our face image extraction model for the infobox image, body image & caption above.

Consider now the processing of an arbitrary Wikipedia biography page of an identity where we find M images of at least a certain size. For each image, we run a face detector, and find N_m faces of some minimum size. We define the faces as $\{\{x_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}$, where, x_{mn} is the n^{th} face from the m^{th} image. For each detected instance of a face, text and meta data are transformed into feature vectors that will be used to determine if the face indeed corresponds to the biography subject. For our text analysis we use information extracted from image file names, image captions and other sources. The location of an image in the page is an example of what we refer to as meta-data. We also treat other information about the image that is not directly involved in facial comparisons as meta-data, ex. the relative size of a face to other faces detected in an image. The bottom layer or set of random variables $\{X\}$ in Figure 1 are used to encode a set of K different text and metadata features for each face. We discuss their nature and the precise definitions of these features in more detail below and in our supplementary material. Each detected face thus has an associated text and metadata feature vector $X_{mn} = [X_1^{(mn)}, X_2^{(mn)}, \cdots X_K^{(mn)}]^T$. These features are used as the input to our model for $P(Y_{mn}|X_{mn})$, where the random variables $\{Y\} = \{\{Y_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}$ are a set of binary target vs. not target indicator variables corresponding to each face, x_{mn}. Inferring these variables jointly corresponds to the goal of our mining model, i.e. finding the faces the correspond to the subject of the biography.

In our example for Mr. Parks, the face detector found a single face in the first image, while in the second image it found three faces. For this specific example, we therefore have three cross image face comparisons that we shall use to aid our disambiguation. The visual similarity of a face pair, $\{x_{mn}, x_{m'n'}\}$, is represented by D_l, where l is an index of all L cross image pairs. Our model for cross image comparisons is encoded withing $p(D_l|Y, Y')$.

Finally, to encode the fact that there is not typically more than one face belonging to the biography subject in a given image we use a constraint variable S_m for each image m. S_m is the child of the indicator variables associated with all the faces of a given image. We then use the corresponding conditional distribution to encode the intuition above as a soft constraint.

With these components defined above, the joint conditional distribution defined by the general case of our model is given by

$$p(\{\{Y_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{D_l\}_{l=1}^{L}, \{S_m\}_{m=1}^{M} | \{\{X_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(Y_{mn}|X_{mn}) p(S_m|\{Y_{mn'}\}_{n'=1}^{N'_m}) \prod_{l=1}^{L} p(D_l|\{Y_{m'_l n'_l}, Y_{m''_l n''_l}\}). \quad (1)$$

Our facial identity resolution problem corresponds to the inference problem of computing the Most Probable Explanation (MPE), Y^* for Y under our model, conditioned on our observations $\{\{\tilde{X}_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{\tilde{D}_l\}_{l=1}^{L}, \{\tilde{S}_m\}_{m=1}^{M}$, corresponding to

$$Y^* = \arg\max_{Y} p(Y|\{\{\tilde{X}_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{\tilde{D}_l\}_{l=1}^{L}, \{\tilde{S}_m\}_{m=1}^{M})$$

As we use a probabilistic formulation we can compute or estimate the probability of any specific assignment to Y using our model. For our facial co-reference experiments, we used a brute force search for the MPE when the number of indicator variables in Y is smaller; while for larger sets of Y, we have developed and use a chunk based resolution protocol discussed below.

Text and image metadata features, $F = \{f_k(X_k^{(mn)}, Y_{mn})\}_{k=1}^{K}$ for each face are used to make predictions in the joint model via a set of discriminative Maximum Entropy Model (MEM) classifiers for each $p(Y_{mn}|X_{mn})$ in the dynamically instantiated model.

Faces are compared across images using fast comparisons between discriminative feature vectors as follows. Given two feature vectors, $\mathbf{x}$ and $\mathbf{y}$, the cosine similarity between $\mathbf{x}$ and $\mathbf{y}$ is simply $CS(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})(\|\mathbf{x}\|\|\mathbf{y}\|)^{-1}$. We model the distributions over cross image face comparisons, or $p(D_l|\{Y_{m'_l n'_l}, Y_{m''_l n''_l}\})$, using discrete distributions for cosine distances between faces. The underlying faces are registered and transformed into feature representations based on local binary patterns as we discuss in more detail below. We use two different discrete distributions as building blocks in the model, one for the distances between faces of the same person, one for the distances between different people. The distributions are obtained by quantizing cosine distances into 20 bins and gathering the appropriate statistics.

However, to further increase the separation of cosine distances between comparisons between same identity vs different identity faces we use variations of the Cosine Similarity Metric Learning (CSML) technique proposed in Nguyen and Bai (2010). This technique allows us to perform linear discriminative dimensionality reduction with a matrix $\mathbf{A}$, based on comparisons of the form

$$CS(\mathbf{x}, \mathbf{y}, \mathbf{A}) = \frac{(\mathbf{Ax})^T(\mathbf{Ay})}{\|\mathbf{Ax}\|\|\mathbf{Ay}\|}. \quad (2)$$

We use this technique as it is fast, and among the leading performers in the Labeled Faces in the Wild (LFW) evaluations in the restricted settings. We discuss this procedure and our particular variation of it in more detail below.

We use these distance distributions associated with comparisons between faces of the same identity vs different identities for modeling the following cases. For an input face pair, $\{x_{m'_l n'_l}, x_{m''_l n''_l}\}$, the corresponding binary labels for their indicator variables, $\{Y_{m'_l n'_l}, Y_{m''_l n''_l}\}$ have four possible configurations: (1) both faces are of the biography subject, (2) the first is, (3) second is the subject, or (4) neither faces are. We model cases (2) and (3) using a single never-same distribution. We model case (4) allowing a small probability that non-subject faces across images are the of the same identity (e.g. spouses, friends, etc.). The same and the never-same distributions over cosine distances are modeled using ($n_s = 3000$) positive and ($n_d = 3000$) negative pairs from the LFW, while the rarely same class is modeled through a weighted combination of positives and negatives with weight parameters w_1 and w_2, estimated using cross validation with a 2D grid search.

It is also important to note that high quality registration of facial images is essential to produce high quality visual comparisons for face verifiers. We therefore discuss the steps for face registration and processing in more detail in the next section.

The binary configuration constraint distribution, $p(S_m|\{Y_{mn}\}_{n=1}^{N_m})$, encodes the fact that it is unlikely that two faces of the same individual appear within the same image. The situation is unlikely but can happen, for example consider the second image in Figure 1 in which there is a second face of Richard Parks in the background which has not been detected due to an occlusion. For a set of faces, $\{x_{mn}\}_{n=1}^{N_m}$, contained within the same image, m, one technique for encoding configuration constraints is through the use of the following conditional distribution for a common child in the network. If none or one of the faces detected in the image belongs to the target identity, we have a normal image (i.e. $S_m = 1$). If two or more faces in the same image belong to the same identity, the constraint of being a normal image is not satisfied. To enforce the constraint during MPE inference we set the observation to $S_m = \tilde{S_m} = 1$, i.e. the constraint is satisfied. Since this type of constraint is usually, but not always satisfied one can captured such a notion using

$$p(\tilde{S}_m|\{Y_{mn}\}_{n=1}^{N_m}) = \begin{cases} q & \text{1 or 0 faces in image of target,} \\ 1-q & \geq \text{2 faces of target,} \end{cases}$$

where q is close but not equal to 1.

To deal with longer sequences, we use a chunk-based approach for ≥ 8 faces. Inference is resolved through chunks of size 7 using a strategy corresponding to a variation of blocked iterated conditional modes (ICM). At each chunk base resolution step, the system is provided with the most probable two faces as pivots from earlier step(s). We initialize the pivots with the most confident two faces from our MEM classifier.

Data Processing, Labeling and Features

We downloaded 214,869 images and their corresponding caption texts from 522,986 Wikipedia living people biography sites. Then, we used the OpenCV face detector (Viola and Jones 2004) to extract faces; for each detection, the faces were cut out from images with an additional 1/3 background to make the data compatible to the LFW benchmark. Roughly one in every three images had at least one face of at least a moderate resolution (40x40 pixels) and we used this as the minimum size for inclusion in our experiments. Among those faces 56.71% were from people with only one face on their biography page. The number of identities that had at least one face is 64,291.

For model evaluations, we sampled and labeled a portion of our data following a stratified sampling approach. More specifically, we grouped and sampled people based on their number of faces. Faces were labeled as true examples of the subject, not examples of the subject or as noisy (photographs, not faces). We randomly selected 250 identities for the most prevalent case where only one face was detected. For identities with ≥ 8 faces, we labeled all faces; while for remaining groups (groups with 2-7 faces), faces from an average 160 identities were labeled.

Text and Metadata Feature Extraction

For person name detections in the caption text, we used the Stanford Named Entity Detector (NED) (Finkel, Grenager, and Manning 2005) and derive various other features from these detections. We have classified the feature definitions of our facial identity resolution model into two general categories: (a) face-pair features, and (b) per-face features. The per-face features are again divided into (i) unigrams: a single and independent feature, and (ii) the logical conjunctions of unigram features which capture first order interaction effects. The local MEMs use all or subsets of the per-face features (based on a specific model setting as described in the experiments section) that defines the feature set, X_{mn} for our models. We also use a set of heuristic comparisons such as relative image size and other meta image features for our text and image models. Below, we provide the definition of a binary feature, *nameInImageFile*, as an example from the larger list of features which is given in Appendix I.

nameInImageFile: This is a binary feature representing whether the person's name appears in the image file name or not. A positive match is defined as if any part (either first name or last name) of the person's name is at least of 3 characters long and a match is found in the image file name.

$$f_k(X_k^{(mn)}, Y_{mn}) = \begin{cases} 1 & \text{if the person's name is found} \\ & \text{in the image file name} \\ 0 & \text{otherwise} \end{cases}$$

Face Registration, Features & Comparisons

High quality visual comparisons are critical for the facial identity resolution problem. Virtually all the top performing methods on the LFW evaluation use commercially aligned faces. To provide the visual comparison part of our model with the best registrations and features possible using an open source documented technique, we have developed our own pose based alignment pipeline. Figure 2 shows the processing steps of our pipeline: an input image is first classified into one of three pose categories (left, center or right facing) using a histogram of gradients + SVM based pose classifier which yields 98.8% accuracy on a 50% test-train split evaluation using the PUT database. We then identify 2-5 spatially consistent keypoints using a variant of the keypoint search algorithm discussed in more detail in Hasan and Pal (2011). These keypoints are then used to align faces to one of three different pose based coordinate frames using a similarity transformation. Our experiments have found that this pipeline yields performance on par with the LFWa commercial alignments.

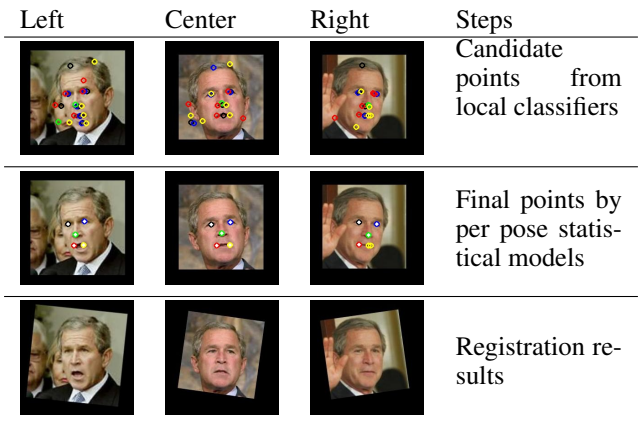

Figure 2: Pose based alignment pipeline steps

When using non-registered faces in our mining experiments, we used the face bounding box area, returned by the OpenCV face detector (Viola and Jones 2004) as the definition of a face. This area is then rescaled to a size of 110x110. For both our mining and verification experiments, when using registered faces we first selected a reference patch of size 80x150 through a reference point, estimated from the locations of the two eyes and the nose tip in the common warping coordinate frame as done in Hasan and Pal (2011). Local Binary Pattern (LBP) features (Ojala, Pietikäinen, and Mäenpää 2001) are then extracted for a non overlapping block size of 10x10.

As we discussed above, we learn discriminative linear projections of these high dimensional feature vectors based on LPB features. For learning we used the LFW view2 dataset and a slight variation of the Cosine Similarity Metric Learning (CSML) technique in Nguyen and Bai (2010). The basic idea is to push the positive and negative samples towards the direction $+1$ and -1 respectively, and thus maximize the between class distance in the cosine space. More specifically, to learn $\mathbf{A}$, from n labeled examples, $\{\mathbf{x}_i, \mathbf{y}_i, l_i\}_{i=1}^n$, where $(\mathbf{x}_i, \mathbf{y}_i)$ is data instance with label $l_i \in \{+1, -1\}$. The model also uses a regularization term to control over fitting based on the L_2 entry-wise distance of the matrix from the whitened PCA solution (i.e. transforming the elements of the matrix into a vector and using the L_2 norm). The complete formulation is based on maximizing

$$f(\mathbf{A}) = \sum_{i \in Pos} CS(\mathbf{x}_i, \mathbf{y}_i, \mathbf{A}) - \alpha \sum_{i \in Neg} CS(\mathbf{x}_i, \mathbf{y}_i, \mathbf{A}) \quad (3)$$

$$- \beta \|\text{vec}(\mathbf{A} - \mathbf{A}_0)\|^2. \quad (4)$$

Equation (4) uses hyper-parameters α, which balances the relative importance given to positive matches vs. negative matches, and β, which controls the strength of the regularization term.

In contrast to the original CSML approach we use a minor modification to the underlying objective function based on one minus the usual cosine distance all squared. We therefore refer to this approach as CSML^2. We also used a few other minor changes to the algorithm to speed it up. Our preliminary experiments indicated that this technique gave a small but not statistically significant boost in performance, but was roughly 50% faster.

In our face mining experiments we used CSML^2 cosine distances learned from the square root LBP features as the underlying discretized observation that was given to the graphical model. In our verification experiments, we used 18 different cosine distances features. These cosine distances are based on the raw and sqrt of : (i) intensity, (ii) LBP, and (iii) Hierarchical LBP (HLBP) features. The HLBP was computed for three levels, starting with the whole image as a patch, and successively dividing into four blocks; then concatenating the feature vectors. A combination of these six feature types for each projection: PCA, Whitened PCA (WPCA), and CSML^2 yield the 18 cosine features. Before learning these CSML^2 projections, the LBP feature vectors were first reduced to 500 dimension through a Principal Component Analysis (PCA) projection. The final CSML^2 projection has 200 dimensions.

Experiments and Analysis

We provide two broad classes of experiments: First, given the importance of high quality face comparisons for identity resolution we provide an evaluation of our face verification techniques using both the widely used LFW evaluation and the face of Wikipedia. We compare our pose guided face verifiers with state of the art verification protocols. In this way we also provide a set of standard baseline verification results for the community using this new Wikipedia based dataset. Second, we provide an extensive set of comparisons of different face mining baselines consisting of different heuristics such as: those using only images and other techniques using only text and meta-data information within

independent classifiers. We then compare different variations of our probabilistic technique which integrates information into dynamically instantiated probabilistic models. Throughout these experiments we examine the impact of alignment on the quality of extraction.

Face Verification in the Wild (LFW & Wikipedia)

Figure 3 compares face verification models using the standard LFW ROC curves. Results are reported for our pose based model on two versions of the LFW data: raw LFW (LFW), and commercially aligned LFW (LFWa). When using the raw LFW or our Wikipedia faces, we aligned images through our pose-based registration pipeline, while for experiments with the LFWa we just used our pose based verification protocol where different SVMs are used for different types of comparisons across poses.

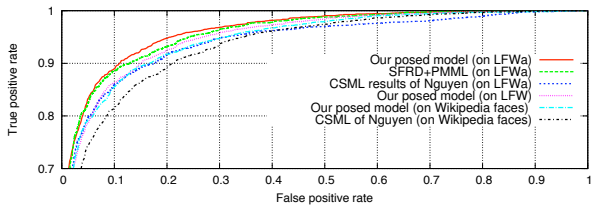

Figure 3: ROC curves for LFW and Wikipedia experiments

We applied our per-pose comparison SVM technique to both the LFWa alignments and our own complete per-pose registration pipeline. Our per pose registration method yields higher performance for comparisons between side profiles of the same orientation (92.4% vs 91.9%), and for side profile comparisons when mirroring is used for opposite orientations (91.5% vs 89.8%), significant under a t-test at the 95% level. However the LFW faces consist of primarily center facing poses, so that across all the comparisons the methods were not statistically different. However, both posed and non-posed registrations yield only $\sim$82% for left-right side profile comparisons without the use of mirroring - which is highly significant, having $p < .001$. Using different SVMs for each type of comparison across poses and mirroring for off center poses we achieve an accuracy of 88.4% using our complete pipeline and 90.0% using the LFWa. Both of these levels of performance would be at the top of the evaluation for verification accuracy for the LFW restricted setting.

Figure 3 also shows the ROC curves for a randomly chosen 3000 positive and 3000 negative Wikipedia face pairs. We use the same 18 LBP features derived from the LFW view2 data as before. We can see that this Wikipedia verification protocol shows a similar performance profile to the LFW evaluation set. While not the main focus of our paper, we see here that our posed based processing techniques can increase performance and in fact yields state of the art performance on the highly competitive LFW evaluation.

Face Mining and Identity Resolution

Table 2 compares mining results using various methods for people with at-least two faces. For each face count group, a randomly chosen 70% of its labeled instances plus all labeled data from its immediate above and below group (if any) were used as training, while the remaining 30% of the examples were used for testing. The results are averaged over 10 runs. We provide aligned and unaligned results if applicable. At the bottom of the table we also give the number of biographies and the % of faces that were indeed of the subject for each group.

First, we provide an image only baseline experiment which follows two simple steps : first find a reference face from someone's Wikipedia page, then using this reference face verify the remaining faces as positives or negatives. The first step follows two ordered heuristic rules: (a) use the first single face image as the reference, and (b) if no reference face is found in a), use the largest face from the first image as the reference. For this image only baseline experiment, we randomly selected 500 positive and 500 negative pairs from faces exclusive to a test group, and learned our CSML^2 verifier for square root LBP features. This heuristic image only approach yielded 61% expected average accuracy with unaligned images and 63% with aligned images.

We also provide a text only baseline classifier that uses independent MEMs for each detected face. The results of a third image-text baseline, and our joint model are also given, which use all modality information available: text, images, and meta-data. The image-text baseline also uses heuristic features derived from comparing images as input to MEM classifiers and yields 71% using aligned faces.

Unsurprisingly, the joint model does not improve dramatically upon the image-text baseline when unaligned faces are used. Since model sub-components are coupled via the quality of the visual comparisons this is to be expected. However, the joint model improves dramatically when aligned faces are used, yielding an expected average accuracy of 77%. The average standard error across these experiments was fairly stable at $\sim$1.2%.

Method	Number of faces detected					Exp.
	2	3	4	5-7	≥ 8	Avg.
Using unaligned faces						
Image only	60	61	58	61	67	61
Text only	69	65	65	62	65	66
Image-text	70	73	71	69	70	71
Joint model	74	72	70	68	71	72
Using aligned faces						
Image only	62	63	61	62	69	63
Image-text	72	74	74	68	72	72
Joint model	**78**	**80**	**77**	**71**	**74**	**77**
# of Bios	7920	2374	1148	1081	468	
% subject	61	53	42	35	29	

Table 2: Accuracies (%) for people with at-least **2** faces.

Among the randomly sampled 250 faces from the group with a single face, 17 (7%) were noisy in the sense that they were either a non face, or a non photograph face (a drawing or a cartoon face), or a face that couldn't be clearly labeled as positive or negative. Out of the 233 photographic faces 231 (99.1%) were true positives, i.e. true instances of our person of interest.

The closest previous work to ours of which we are aware is the "Names and faces in the News" work of Berg et al. (2004b). While the differences of their setup make a direct comparison of methods impossible we discuss their work here to give some additional context to our results. In their work, 1,000 faces were randomly selected from their 45,000 face database, and were hand labeled with person names for model evaluations. Their images were taken from press photos containing small numbers of faces per image. Performance evaluations were conducted using an independent language model (no appearance model), and on a combined appearance and language model. They have reported their best name-face association accuracies for the following four setups: (i) A language model with Expectation Maximization (EM) training: 56%, (ii) Language model with maximal assignment clustering (MM): 67%, (iii) A context understanding joint model (Naive Bayes language model + appearance model): 77%, and (iii) A context understanding joint model (Maximum Entropy language model + appearance model): 78%.

Final Discussion and Conclusions

Comparing the LFW with the Faces of Wikipedia we believe the slight reductions in verification and recognition performance are due in large part to greater age variability on Wikipedia. In terms of extraction performance, our preliminary error analysis indicates that the majority of errors are caused by subject faces that were not detected. We therefore plan to use a higher quality detector for the next release of the dataset in which we are also using Mechanical Turk to scale up labeling.

In summary, we have made a number of contributions in our work here. First and foremost, we have shown how to construct well-defined probabilistic models that formally account for the uncertainty arising from the analysis of both face comparisons and many different natural language statements concerning the content of different images found throughout a document. Along the way to this primary contribution we have developed and use a number of registration and verification pipelines that yield state of the art performance on the LFW face verification benchmark. Finally, we release the Faces of Wikipedia database along with these experiments to the community. We believe the data and these experiments will be of great use to the community, allowing verification, recognition and visual information extraction experiments to be performed and compared at unprecedented scales in terms of the number of images and identities.

Acknowledgements

We thank Erik Learned-Miller and Yoshua Bengio for discussion on various aspects of this work. This work was funded in part by a Google Research award, and the Natural Science and Engineering Research Council (NSERC) Discovery Grants Program. We sincerely thank these sponsors.

Appendix I: Text and Metadata Features

The features, $\{f_k(X_k^{(mn)}, Y_{mn})\}_{k=1}^{K}$, used as input to the per-face classifiers consist of both the primary features defined below and composite features composed from the logical conjunction of certain pairs of binary features. The primary feature set consists of the following.

nameInImageFile: A binary feature representing whether the person's name appears in the image file name or not. A positive match is defined as if any part (either first name or last name) of the person's name is at least of 3 characters long and a match is found in the image file name.

posWordInFname : A binary feature representing whether there appears any positive word in the image file name. A word is considered to be positive if it provides evidence for a face to be positive. For example, if there appears a word, 'portrait', it provides clues that the detected face is a portrait of our person of interest. The positive words are extracted from caption texts and image file names of positive faces. In file names, we manually searched for positive words, where for caption texts the top listed (high frequency) words, excluding the stop words and the Named Entities (NE) are defined as positive words. Some examples of these positive words are: crop, portrait, address, pose, speak, waves, delivers, honored, taken, and poster.

negWordInFname : A binary feature representing whether there appears any negative word in the image file name. A word is considered as negative if it induces noise for a face to be positive. For example, the word 'and' indicates that there might appear a second person in the image. Usually, the conjunct words, like 'and', and 'with', and relationship words, like, mother, spouse are examples of such words. Negative words were extracted from file names of images where true negative faces were found. Some examples of such negative words include: puppet, partner, father, mother, wife, spouse, son, daughter, and brother.

psNameInCaption : A binary feature for whether the person's name appeared in the caption text or not. A positive match is defined as if any part (either first name or last name) of the person's name is at least of 3 characters long and a match is found with the person names, returned by a Named Entity Detector (NED), for an input caption text.

secondNameInCaption : A binary feature representing whether any second person's name (other than our person of interest) is detected in the caption text.

posWordInCaption : A binary feature representing whether there appears any positive word in the caption text. The definition of a positive word here is similar to our previous definition for *posWordInFname*.

negWordInCaption : A binary feature representing whether there appears any negative word in the caption text or not.

leftWordOne, and **leftWordTwo** : A *left-word* is a linguistic token that generally appears left to a person name for whom we have a positive face. These two binary features, *leftWordOne*, and *leftWordTwo* represent whether there appears any *left-word* within the immediate left two positions of the person name being detected by the NED (if any). The *left-word* list is extracted from labeled training examples.

rightWordOne,**rightWordTwo** : These two binary features represent whether there appears any *right-word* within the immediate two right positions of the person name being detected by the NED (if any). The *right-word* is defined following a similar principle as the *left-word*.
pr_imSource : A binary feature that encodes if the parent image is from an Infobox of the Wikipedia page.
pr_imNumOfFaces : A discrete feature with five possible integer values, from 0 to 4, representing the number of faces, detected in an image.
isTheLargestFace : A binary feature representing whether the face is the largest among all its siblings.
theClosestMatch : For a face, x_{mn}, this feature encodes the bin index of its closest visual similarity match from all cross-image pairs, $\{D_l\}_l^L$. We use the cross-image pair definition, D_l, as discussed in the main manuscript. We discretized the square root LBP CSML^2 visual similarity distances into 5 bins.

We used the following feature conjunctions: posWordInFName $\wedge$ negWordInImageFile, posWordInFName $\wedge$ nameInImageFile, posWordInFName $\wedge$ isTheLargestFace, negWordInImageFile $\wedge$ nameInImageFile, negWordInImageFile $\wedge$ isTheLargestFace, and nameInImageFile $\wedge$ isTheLargestFace.

References

Angelova, A.; Abu-Mostafam, Y.; and Perona, P. 2005. Pruning training sets for learning of object categories. In *CVPR*, 494–501.

Anguelov, D.; Lee, K.; Gokturk, S. B.; and Sumengen, B. 2007. Contextual identity recognition in personal photo albums. In *CVPR*, 1–7.

Berg, T. L.; Berg, E. C.; Edwards, J.; Maire, M.; White, R.; Teh, Y.; Learned-Miller, E.; and Forsyth, D. A. 2004a. Names and faces. Technical report.

Berg, T. L.; Berg, E. C.; Edwards, J.; Maire, M.; White, R.; Teh, Y.; Learned-Miller, E.; and Forsyth, D. A. 2004b. Names and faces in the news. In *CVPR*, 848–854.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. ACL*, 363–370.

Guillaumin, M.; Mensink, T.; Verbeek, J.; and Schmid, C. 2012. Face recognition from caption-based supervision. *International Journal of Computer Vision (IJCV)* 96(1):64–82.

Hasan, M., and Pal, C. 2011. Improving the alignment of faces for recognition. In *IEEE ROSE Symposium*, 249–254.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Kittur, A.; Chi, E.; and Suh, B. 2009. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *CHI*, 1509–1512.

Kumar, N.; Belhumeur, P.; and Nayar, S. 2008. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 340–353.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *ICCV*.

Nguyen, H. V., and Bai, L. 2010. Cosine similarity metric learning for face verification. In *ACCV*, 709–720.

Ojala, T.; Pietikäinen, M.; and Mäenpää, T. 2001. A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. *CAPR* (2013):397–406.

Susskind, J.; Anderson, A.; and Hinton, G. 2010. The toronto face database. Technical report, University of Toronto.

Torralba, A.; Fergus, R.; and Freeman, W. T. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IJCV* 30(11):1958–1970.

Viola, P., and Jones, M. J. 2004. Robust real-time face detection. *IJCV* 57(2):137–154.

Wolf, L.; Hassner, T.; and Maoz, I. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 529–534.

Wong, Y.; Chen, S.; Mau, S.; Sanderson, C.; and Lovell, B. C. 2011. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *CVPR 2001 Workshop*.

Zhang, L.; Hu, Y.; Li, M.; Ma, W.; and Zhang, H. 2004. Efficient propagation for face annotation in family albums. In *Proceedings of the 12th annual ACM international conference on Multimedia*, 716–723. ACM.